

---

# Linearized Alternating Direction Method with Adaptive Penalty for Low-Rank Representation

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

Low-rank representation (LRR) is an effective method for subspace clustering and has found wide applications in computer vision and machine learning. The existing LRR solver is based on the alternating direction method (ADM). It suffers from  $O(n^3)$  computation complexity due to the matrix-matrix multiplications and matrix inversions, even if partial SVD is used. Moreover, introducing auxiliary variables also slows down the convergence. Such a heavy computation load prevents LRR from large scale applications. In this paper, we generalize ADM by linearizing the quadratic penalty term and allowing the penalty to change adaptively. We also propose a novel rule to update the penalty such that the convergence is fast. With our linearized ADM with adaptive penalty (LADMAP) method, it is unnecessary to introduce auxiliary variables and invert matrices. The matrix-matrix multiplications are further alleviated by using the skinny SVD representation technique. As a result, we arrive at an algorithm for LRR with complexity  $O(rn^2)$ , where  $r$  is the rank of the representation matrix. Numerical experiments verify that for LRR our LADMAP method is much faster than state-of-the-art algorithms. Although we only present the results on LRR, LADMAP actually can be applied to solving more general convex programs.

## 1 Introduction

Recently, compressive sensing [5] and sparse representation [20] have been hot research topics and also have found abundant applications in signal processing and machine learning. Many of the problems in these fields can be formulated as the following convex programs:

$$\min_{\mathbf{x}, \mathbf{y}} f(\mathbf{x}) + g(\mathbf{y}), \text{ s.t. } \mathcal{A}(\mathbf{x}) + \mathcal{B}(\mathbf{y}) = \mathbf{c}, \quad (1)$$

where  $\mathbf{x}$ ,  $\mathbf{y}$  and  $\mathbf{c}$  could be either vectors or matrices,  $f$  and  $g$  are convex functions (e.g., the nuclear norm  $\|\cdot\|_*$  [2], Frobenius norm  $\|\cdot\|$ ,  $l_{2,1}$  norm  $\|\cdot\|_{2,1}$  [14], and  $l_1$  norm  $\|\cdot\|_1$ ), and  $\mathcal{A}$  and  $\mathcal{B}$  are linear mappings.

As a measure of 2D sparsity, nuclear norm minimization (namely,  $f(\mathbf{X}) = \|\mathbf{X}\|_*$ ) has now attracted a lot of attention and has been successfully applied to video processing [10], matrix recovery [4], unsupervised learning [13] and semi-supervised learning [8]. Typical problems are matrix completion [4], robust principal component analysis [19] and their combination [3]. A nuclear norm minimization problem could be reformulated as a semidefinite program [4], hence could be solved by any off-the-shelf interior point based toolbox, such as CVX. However, interior point methods cannot handle large scale matrices due to their  $O(n^6)$  complexity in each iteration, where  $n \times n$  is the matrix size. To overcome this issue, several first-order algorithms have been developed to solve nuclear norm minimization problems. One method is the singular value thresholding (SVT) algorithm [2] which applies the soft-thresholding operator to the singular values of a certain matrix in

each iteration. The accelerated proximal gradient (APG) algorithm [17] is also a popular choice due to its guaranteed  $O(k^{-2})$  convergence rate, where  $k$  is the iteration number. The alternating direction method (ADM) has also regained a lot of attention [12, 16]. It updates the variables alternately by minimizing the augmented Lagrangian function with respect to the variables in a Gauss-Seidel manner.

In 2010, Liu et al. [13] proposed the low-rank representation (LRR) for robust subspace clustering. Unlike the sparse representation [6], which minimizes the number of nonzero entries in the representation matrix, LRR seeks to minimize the rank of the representation matrix. The mathematical model of LRR is as follows<sup>1</sup>:

$$\min_{\mathbf{Z}, \mathbf{E}} \|\mathbf{Z}\|_* + \mu \|\mathbf{E}\|_{2,1}, \text{ s.t. } \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \quad (2)$$

where  $\mathbf{X}$  is the data matrix. LRR has found wide applications in computer vision and machine learning, e.g., motion segmentation, face clustering, and temporal segmentation [13, 15, 7].

The existing LRR solver [13] is based on ADM. It suffers from  $O(n^3)$  computation complexity due to the matrix-matrix multiplications and matrix inversions. Moreover, introducing auxiliary variables also slows down the convergence, as there are more variables and constraints. Such a heavy computation load prevents LRR from large scale applications. In this paper, we generalize ADM by linearizing the quadratic penalty term and allowing the penalty to change adaptively. Linearization makes the auxiliary variables unnecessary, hence waiving the matrix inversions, while variable penalty makes the convergence fast. We also propose a novel and simple rule to update the penalty. We prove the global convergence of linearized ADM with adaptive penalty (LADMAP) and apply it to LRR, obtaining faster convergence speed than the original solver. By further representing  $\mathbf{Z}$  as its skinny SVD and utilizing an advanced functionality of the PROPACK [11] package, the complexity of solving LRR by LADMAP becomes only  $O(rn^2)$ , as there is no full sized matrix-matrix multiplications, where  $r$  is the rank of the optimal  $\mathbf{Z}$ . Although we only present the numerical results on LRR, LADMAP actually can be applied to solving more general convex programs.

Our work is inspired by Yang et al. [21]. Nonetheless, the difference of our work from theirs is distinct. First, they only proved the convergence of linearized ADM (LADM) for a specific problem, namely nuclear norm regularization. Their proof utilized some special properties of the nuclear norm, while we prove the convergence of LADM for general problems in (1). Second, they only proved in the case of fixed penalty, while we prove in the case of variable penalty. Although they mentioned the dynamic updating rule proposed in [9], their proof cannot be straightforwardly applied to the case of variable penalty. Moreover, that rule is for ADM only. Third, the convergence speed of LADM heavily depends on the choice of penalty. So it is difficult to choose an optimal fixed penalty that fits for different problems and problem sizes, while our novel updating rule for the penalty, although simple, is effective for different problems and problem sizes.

## 2 Linearized Alternating Direction Method with Adaptive Penalty

### 2.1 The Alternating Direction Method

ADM is now very popular in solving large scale sparse representation problems [1]. When solving (1) by ADM, one operates on the following augmented Lagrangian function:

$$\mathcal{L}_A(\mathbf{x}, \mathbf{y}, \lambda) = f(\mathbf{x}) + g(\mathbf{y}) + \langle \lambda, \mathcal{A}(\mathbf{x}) + \mathcal{B}(\mathbf{y}) - \mathbf{y} \rangle + \frac{\beta}{2} \|\mathcal{A}(\mathbf{x}) + \mathcal{B}(\mathbf{y}) - \mathbf{c}\|^2, \quad (3)$$

where  $\lambda$  is the Lagrange multiplier,  $\langle \cdot, \cdot \rangle$  is the inner product, and  $\beta > 0$  is the penalty parameter. The usual augmented Lagrange multiplier method is to minimize  $\mathcal{L}_A$  w.r.t.  $\mathbf{x}$  and  $\mathbf{y}$  simultaneously. This is usually difficult and does not exploit the fact that the objective function is separable. To remedy this issue, ADM decomposes the minimization of  $\mathcal{L}_A$  w.r.t.  $(\mathbf{x}, \mathbf{y})$  into two subproblems that minimize w.r.t.  $\mathbf{x}$  and  $\mathbf{y}$ , respectively. More specifically, the iterations of ADM go as follows:

$$\mathbf{x}_{k+1} = \arg \min_{\mathbf{x}} \mathcal{L}_A(\mathbf{x}, \mathbf{y}_k, \lambda_k), \quad (4)$$

$$\mathbf{y}_{k+1} = \arg \min_{\mathbf{y}} \mathcal{L}_A(\mathbf{x}_{k+1}, \mathbf{y}, \lambda_k), \quad (5)$$

$$\lambda_{k+1} = \lambda_k + \beta[\mathcal{A}(\mathbf{x}_{k+1}) + \mathcal{B}(\mathbf{y}_{k+1}) - \mathbf{c}]. \quad (6)$$

<sup>1</sup>Here we switch to bold capital letters in order to emphasize that the variables are matrices.

In compressive sensing and sparse representation, as  $f$  and  $g$  are usually matrix or vector norms, the subproblems (4) and (5) usually have closed form solutions when  $\mathcal{A}$  and  $\mathcal{B}$  are identities [2, 13, 22]. In this case, ADM is appealing. However, in many problems  $\mathcal{A}$  and  $\mathcal{B}$  are not identities. For example, in matrix completion  $\mathcal{A}$  can be a selection matrix, and in 1D sparse representation  $\mathcal{A}$  can be a general matrix. In this case, there are no closed form solutions to (4) and (5). To overcome this difficulty, a common strategy is to introduce auxiliary variables [13, 1]  $\mathbf{u}$  and  $\mathbf{v}$  and reformulate problem (1) into an equivalent one:

$$\min_{\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v}} f(\mathbf{x}) + g(\mathbf{y}), \text{ s.t. } \mathcal{A}(\mathbf{u}) + \mathcal{B}(\mathbf{v}) = \mathbf{c}, \mathbf{x} = \mathbf{u}, \mathbf{y} = \mathbf{v}, \quad (7)$$

and the corresponding ADM iterations analogous to (4)-(6) can be deduced. With more variables and more constraints, the convergence of ADM becomes slower. Moreover, to update  $\mathbf{u}$  and  $\mathbf{v}$ , whose subproblems are least squares problems, matrix inversions are often necessary.

## 2.2 Linearized ADM

To avoid introducing auxiliary variables and still solve subproblems (4) and (5) efficiently, inspired by Yang et al. [21], we propose a linearization technique for (4) and (5). To further accelerate the convergence of the algorithm, we also propose an adaptive rule for updating the penalty parameter.

With minor algebra, one can see that subproblem (4) is equivalent to

$$\mathbf{x}_{k+1} = \arg \min_{\mathbf{x}} f(\mathbf{x}) + \frac{\beta}{2} \|\mathcal{A}(\mathbf{x}) + \mathcal{B}(\mathbf{y}_k) - \mathbf{c} + \lambda_k / \beta\|^2. \quad (8)$$

By linearizing the quadratic term in (8) at  $\mathbf{x}_k$  and adding a proximal term, we have the following approximation:

$$\begin{aligned} \mathbf{x}_{k+1} &= \arg \min_{\mathbf{x}} f(\mathbf{x}) + \langle \mathcal{A}^*(\lambda_k) + \beta \mathcal{A}^*(\mathcal{A}(\mathbf{x}_k) + \mathcal{B}(\mathbf{y}_k) - \mathbf{c}), \mathbf{x} - \mathbf{x}_k \rangle + \frac{\beta \eta_A}{2} \|\mathbf{x} - \mathbf{x}_k\|^2 \\ &= \arg \min_{\mathbf{x}} f(\mathbf{x}) + \frac{\beta \eta_A}{2} \|\mathbf{x} - \mathbf{x}_k + \mathcal{A}^*(\lambda_k + \beta(\mathcal{A}(\mathbf{x}_k) + \mathcal{B}(\mathbf{y}_k) - \mathbf{c})) / (\beta \eta_A)\|^2, \end{aligned} \quad (9)$$

where  $\mathcal{A}^*$  is the adjoint of  $\mathcal{A}$  and  $\eta_A > 0$  is a parameter whose proper value will be analyzed later. The above approximation resembles that of APG [17], but we do not use APG to solve (4) iteratively.

Similarly, subproblem (5) can be approximated by

$$\mathbf{y}_{k+1} = \arg \min_{\mathbf{y}} g(\mathbf{y}) + \frac{\beta \eta_B}{2} \|\mathbf{y} - \mathbf{y}_k + \mathcal{B}^*(\lambda_k + \beta(\mathcal{A}(\mathbf{x}_{k+1}) + \mathcal{B}(\mathbf{y}_k) - \mathbf{c})) / (\beta \eta_B)\|^2. \quad (10)$$

The update of Lagrange multiplier still goes as (6)<sup>2</sup>.

## 2.3 Adaptive Penalty

In previous ADM and LADM approaches [16, 22, 21], the penalty parameter  $\beta$  is fixed. Some scholars have observed that ADM with a fixed  $\beta$  can converge very slowly and it is nontrivial to choose an optimal fixed  $\beta$ . So is LADM. Thus a dynamic  $\beta$  is preferred in real applications. Although Tao et al. [16] and Yang et al. [21] mentioned He et al.'s adaptive updating rule [9] in their papers, the rule is for ADM only. We propose the following adaptive updating strategy for the penalty parameter  $\beta$ :

$$\beta_{k+1} = \min(\beta_{\max}, \rho \beta_k), \quad (11)$$

where  $\beta_{\max}$  is an upper bound of  $\{\beta_k\}$ . The value of  $\rho$  is defined as

$$\rho = \begin{cases} \rho_0, & \text{if } \beta_k \max(\sqrt{\eta_A} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|, \sqrt{\eta_B} \|\mathbf{y}_{k+1} - \mathbf{y}_k\|) / \|\mathbf{c}\| < \varepsilon_2, \\ 1, & \text{otherwise,} \end{cases} \quad (12)$$

where  $\rho_0 \geq 1$  is a constant. The condition to assign  $\rho = \rho_0$  comes from the analysis on the stopping criteria (see Section 2.5). We recommend that  $\beta_0 = \alpha \varepsilon_2$ , where  $\alpha$  depends on the size of  $\mathbf{c}$ . Our updating rule is fundamentally different from He et al.'s for ADM [9], which aims at balancing the errors in the stopping criteria and involves several parameters.

<sup>2</sup>As in [21], we can also introduce a parameter  $\gamma$  and update  $\lambda$  as  $\lambda_{k+1} = \lambda_k + \gamma \beta [\mathcal{A}(\mathbf{x}_{k+1}) + \mathcal{B}(\mathbf{y}_{k+1}) - \mathbf{c}]$ . We choose not to do so in this paper in order not to make the exposition of LADMAP too complex. The reviewers can refer to Supplementary Material for full details.

## 2.4 Convergence of LADMAP

To prove the convergence of LADMAP, we first have the following propositions.

### Proposition 1

$$-\beta_k \eta_A (\mathbf{x}_{k+1} - \mathbf{x}_k) - \mathcal{A}^*(\tilde{\lambda}_{k+1}) \in \partial f(\mathbf{x}_{k+1}), \quad -\beta_k \eta_B (\mathbf{y}_{k+1} - \mathbf{y}_k) - \mathcal{B}^*(\hat{\lambda}_{k+1}) \in \partial g(\mathbf{y}_{k+1}), \quad (13)$$

where  $\tilde{\lambda}_{k+1} = \lambda_k + \beta_k [\mathcal{A}(\mathbf{x}_k) + \mathcal{B}(\mathbf{y}_k) - \mathbf{c}]$ ,  $\hat{\lambda}_{k+1} = \lambda_k + \beta_k [\mathcal{A}(\mathbf{x}_{k+1}) + \mathcal{B}(\mathbf{y}_k) - \mathbf{c}]$ , and  $\partial f$  and  $\partial g$  are subgradients of  $f$  and  $g$ , respectively.

This can be easily proved by checking the optimality conditions of (9) and (10).

**Proposition 2** Denote the operator norms of  $\mathcal{A}$  and  $\mathcal{B}$  as  $\|\mathcal{A}\|$  and  $\|\mathcal{B}\|$ , respectively. If  $\{\beta_k\}$  is non-decreasing and upper bounded,  $\eta_A > \|\mathcal{A}\|^2$ ,  $\eta_B > \|\mathcal{B}\|^2$ , and  $(x^*, y^*, \lambda^*)$  is any Karush-Kuhn-Tucker (KKT) point of problem (1) (see (14)-(15)), then: (1).  $\{\eta_A \|\mathbf{x}_k - \mathbf{x}^*\|^2 - \|\mathcal{A}(\mathbf{x}_k - \mathbf{x}^*)\|^2 + \eta_B \|\mathbf{y}_k - \mathbf{y}^*\|^2 + \beta_k^{-2} \|\lambda_k - \lambda^*\|^2\}$  is non-increasing. (2).  $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \rightarrow 0$ ,  $\|\mathbf{y}_{k+1} - \mathbf{y}_k\| \rightarrow 0$ ,  $\|\lambda_{k+1} - \lambda_k\| \rightarrow 0$ .

The proof can be found in Supplementary Material. Then we can prove the convergence of LADMAP, as stated in the following theorem.

**Theorem 3** If  $\{\beta_k\}$  is non-decreasing and upper bounded,  $\eta_A > \|\mathcal{A}\|^2$ , and  $\eta_B > \|\mathcal{B}\|^2$ , then the sequence  $\{(\mathbf{x}_k, \mathbf{y}_k, \lambda_k)\}$  generated by LADMAP converges to a KKT point of problem (1).

The proof can be found in Appendix A.

## 2.5 Stopping Criteria

The KKT conditions of problem (1) are that there exists a triple  $(\mathbf{x}^*, \mathbf{y}^*, \lambda^*)$  such that

$$\mathcal{A}(\mathbf{x}^*) + \mathcal{B}(\mathbf{y}^*) - \mathbf{c} = \mathbf{0}, \quad (14)$$

$$-\mathcal{A}^*(\lambda^*) \in \partial f(\mathbf{x}^*), \quad -\mathcal{B}^*(\lambda^*) \in \partial g(\mathbf{y}^*). \quad (15)$$

The triple  $(\mathbf{x}^*, \mathbf{y}^*, \lambda^*)$  is called a KKT point. So the first stopping criterion is the feasibility:

$$\|\mathcal{A}(\mathbf{x}_{k+1}) + \mathcal{B}(\mathbf{y}_{k+1}) - \mathbf{c}\| / \|\mathbf{c}\| < \varepsilon_1. \quad (16)$$

As for the second KKT condition, we rewrite the second part of Proposition 1 as follows

$$-\beta_k [\eta_B (\mathbf{y}_{k+1} - \mathbf{y}_k) + \mathcal{B}^*(\mathcal{A}(\mathbf{x}_{k+1} - \mathbf{x}_k))] - \mathcal{B}^*(\tilde{\lambda}_{k+1}) \in \partial g(\mathbf{y}_{k+1}). \quad (17)$$

So for  $\tilde{\lambda}_{k+1}$  to satisfy the second KKT condition, both  $\beta_k \eta_A \|\mathbf{x}_{k+1} - \mathbf{x}_k\|$  and  $\beta_k \|\eta_B (\mathbf{y}_{k+1} - \mathbf{y}_k) + \mathcal{B}^*(\mathcal{A}(\mathbf{x}_{k+1} - \mathbf{x}_k))\|$  should be small enough. This leads to the second stopping criterion:

$$\beta_k \max(\eta_A \|\mathbf{x}_{k+1} - \mathbf{x}_k\| / \|\mathcal{A}^*(\mathbf{c})\|, \eta_B \|\mathbf{y}_{k+1} - \mathbf{y}_k\| / \|\mathcal{B}^*(\mathbf{c})\|) \leq \varepsilon'_2. \quad (18)$$

By estimating  $\|\mathcal{A}^*(\mathbf{c})\|$  and  $\|\mathcal{B}^*(\mathbf{c})\|$  by  $\sqrt{\eta_A} \|\mathbf{c}\|$  and  $\sqrt{\eta_B} \|\mathbf{c}\|$ , respectively, we arrive at the second stopping criterion in use:

$$\beta_k \max(\sqrt{\eta_A} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|, \sqrt{\eta_B} \|\mathbf{y}_{k+1} - \mathbf{y}_k\|) / \|\mathbf{c}\| \leq \varepsilon_2. \quad (19)$$

Finally, we summarize our LADMAP algorithm in Algorithm 1.

## 3 Applying LADMAP to LRR

In this section, we apply LADMAP to solving the LRR problem (2). We also compare LADMAP with other state-of-the-art algorithms for LRR. The reason we choose LRR as an example of applications of LADMAP is twofold. First, LRR has become an important mathematical model in machine learning. Second, unlike other established nuclear norm minimization problems, such as matrix completion [4] and robust principal component analysis [3], if not carefully treated, the complexity of solving LRR is still  $O(n^3)$ , even if partial SVD is used.

---

**Algorithm 1** LADMAP for Problem (1)

---

**Initialize:** Set  $\varepsilon_1 > 0, \varepsilon_2 > 0, \beta_{\max} \gg \beta_0 > 0, \eta_A > \|\mathcal{A}\|^2, \eta_B > \|\mathcal{B}\|^2, \mathbf{x}_0, \mathbf{y}_0, \lambda_0$ , and  $k \leftarrow 0$ .

**while** (16) or (19) is not satisfied **do**

**Step 1:** Update  $\mathbf{x}$  by solving (9).

**Step 2:** Update  $\mathbf{y}$  by solving (10).

**Step 3:** Update  $\lambda$  by (6).

**Step 4:** Update  $\beta$  by (11) and (12).

**Step 5:**  $k \leftarrow k + 1$ .

**end while**

---

### 3.1 Solving LRR by LADMAP

As the LRR problem (2) is a special case of problem (1), LADMAP can be directly applied to it. The two subproblems both have closed form solutions. In the subproblem for updating  $\mathbf{E}$ , one may apply the  $l_{2,1}$ -norm shrinkage operator [13], with a threshold  $\beta_k^{-1}$ , to matrix  $\mathbf{M}_k = -\mathbf{X}\mathbf{Z}_k + \mathbf{X} - \Lambda_k/\beta_k$ . In the subproblem for updating  $\mathbf{Z}$ , one has to apply the singular value shrinkage operator [2], with a threshold  $(\beta_k\eta_X)^{-1}$ , to matrix  $\mathbf{N}_k = \mathbf{Z}_k - \eta_X^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{Z}_k + \mathbf{E}_{k+1} - \mathbf{X} + \Lambda_k/\beta_k)$ , where  $\eta_X > \sigma_{\max}^2(\mathbf{X})$ . If  $\mathbf{N}_k$  is formed explicitly, the usual technique of partial SVD, using PROPACK [11], with rank prediction [12, 21, 17] can be utilized to compute the leading  $r$  singular values and associated vectors of  $\mathbf{N}_k$  efficiently, making the complexity of SVD computation  $O(rn^2)$ , where  $r$  is the predicted rank of  $\mathbf{Z}_{k+1}$  and  $n$  is the column number of  $\mathbf{X}$ . Note that as  $\beta_k$  is non-decreasing, the predicted rank is almost non-decreasing, making the iterations computationally efficient.

Up to now, LADMAP for LRR is still of complexity  $O(n^3)$ , although partial SVD is already used. This is because forming  $\mathbf{M}_k$  and  $\mathbf{N}_k$  requires full sized matrix-matrix multiplications, e.g.,  $\mathbf{X}\mathbf{Z}_k$ . To break this complexity bound, we introduce another technique to further accelerate LADMAP for LRR. By representing  $\mathbf{Z}_k$  as its skinny SVD:  $\mathbf{Z}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T$ , some of the full sized matrix-matrix multiplications are gone: they are replaced by successive reduced sized matrix-matrix multiplications. For example, when updating  $\mathbf{E}$ ,  $\mathbf{X}\mathbf{Z}_k$  is computed as  $((\mathbf{X}\mathbf{U}_k)\mathbf{\Sigma}_k)\mathbf{V}_k^T$ , reducing the complexity to  $O(rn^2)$ . When computing the partial SVD of  $\mathbf{N}_k$ , things are more complicated. If we form  $\mathbf{N}_k$  explicitly, we will face with computing  $\mathbf{X}^T(\mathbf{X} + \Lambda_k/\beta_k)$ , which is neither low-rank nor sparse<sup>3</sup>. Fortunately, in PROPACK the bi-diagonalizing process of  $\mathbf{N}_k$  is done by the Lanczos procedure [11], which only requires to compute matrix-vector multiplications  $\mathbf{N}_k\mathbf{v}$  and  $\mathbf{u}^T\mathbf{N}_k$ , where  $\mathbf{u}$  and  $\mathbf{v}$  are some vectors in the Lanczos procedure. So we may compute  $\mathbf{N}_k\mathbf{v}$  and  $\mathbf{u}^T\mathbf{N}_k$  by multiplying the vectors  $\mathbf{u}$  and  $\mathbf{v}$  successively with the component matrices in  $\mathbf{N}_k$ , rather than forming  $\mathbf{N}_k$  explicitly. So the computation complexity of partial SVD of  $\mathbf{N}_k$  is still  $O(rn^2)$ . Consequently, with our acceleration techniques, the complexity of our accelerated LADMAP for LRR is  $O(rn^2)$ . The accelerated LADMAP is summarized in Algorithm 2.

### 3.2 Comparison with Other Methods

As shown in [13], the LRR problem can be solved by the classic ADM. However, their algorithm requires an auxiliary variable, matrix-matrix multiplication and inversion of matrices, resulting in  $O(n^3)$  computation complexity and slow convergence.

The LRR problem can also be solved *approximately* by being reformulated to the following unconstrained optimization problem:  $\min_{\mathbf{Z}, \mathbf{E}} \beta(\|\mathbf{Z}\|_* + \mu\|\mathbf{E}\|_{2,1}) + \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{R} - \mathbf{E}\|^2$ , where  $\beta > 0$  is a relaxation parameter. Then APG with the continuation technique [17], which is to reduce  $\beta$  gradually by  $\beta_{k+1} = \max(\beta_{\min}, \theta\beta_k)$ , can be applied to solve this problem<sup>4</sup>. Compared with APG, which can only find an approximate solution, LADMAP can produce a much more accurate solution as it is proven to converge to an exact solution.

---

<sup>3</sup>When forming  $\mathbf{N}_k$  explicitly,  $\mathbf{X}^T\mathbf{X}\mathbf{Z}_k$  can be computed as  $((\mathbf{X}^T(\mathbf{X}\mathbf{U}_k))\mathbf{\Sigma}_k)\mathbf{V}_k^T$ , whose complexity is still  $O(rn^2)$ , while  $\mathbf{X}^T\mathbf{E}_{k+1}$  could also be accelerated as  $\mathbf{E}_{k+1}$  is a column-sparse matrix.

<sup>4</sup>Please see Supplementary Material for the detail of solving LRR by APG.

---

**Algorithm 2** Accelerated LADMAP for LRR (2)

---

**Input:** Observation matrix  $\mathbf{X}$  and parameter  $\mu > 0$ .  
**Initialize:** Set  $\mathbf{E}_0$ ,  $\mathbf{Z}_0$  and  $\mathbf{\Lambda}_0$  to zero matrices, where  $\mathbf{Z}_0$  is represented as  $(\mathbf{U}_0, \mathbf{\Sigma}_0, \mathbf{V}_0) \leftarrow (0, 0, 0)$ . Set  $\varepsilon_1 > 0$ ,  $\varepsilon_2 > 0$ ,  $\beta_{\max} \gg \beta_0 > 0$ ,  $\eta_X > \sigma_{\max}^2(\mathbf{X})$ , and  $k \leftarrow 0$ .  
**while** (16) or (19) is not satisfied **do**  
    **Step 1:** Update  $\mathbf{E}_{k+1} = \arg \min_{\mathbf{E}} \mu \|\mathbf{E}\|_{2,1} + \frac{\beta_k}{2} \|\mathbf{E} + (\mathbf{X}\mathbf{U}_k)\mathbf{\Sigma}_k\mathbf{V}_k^T - \mathbf{X} + \mathbf{\Lambda}_k/\beta_k\|^2$ . This subproblem can be solved by using Lemma 3.2 in [13].  
    **Step 2:** Update the skinny SVD  $(\mathbf{U}_{k+1}, \mathbf{\Sigma}_{k+1}, \mathbf{V}_{k+1})$  of  $\mathbf{Z}_{k+1}$ . First, compute the partial SVD  $\tilde{\mathbf{U}}_r \tilde{\mathbf{\Sigma}}_r \tilde{\mathbf{V}}_r^T$  of the *implicit* matrix  $\mathbf{N}_k$ , which is bi-diagonalized by the successive matrix-vector multiplication technique described in Section 3.1, and the rank  $r$  is predicted as in [12, 21, 17]. Second,  $\mathbf{U}_{k+1} = \tilde{\mathbf{U}}_r(:, 1:r')$ ,  $\mathbf{\Sigma}_{k+1} = \tilde{\mathbf{\Sigma}}_r(1:r', 1:r') - (\beta_k \eta_X)^{-1} \mathbf{I}$ ,  $\mathbf{V}_{k+1} = \tilde{\mathbf{V}}_r(:, 1:r')$ , where  $r'$  is the number of singular values in  $\tilde{\mathbf{\Sigma}}_r$  that are greater than  $(\beta_k \eta_X)^{-1}$ .  
    **Step 3:** Update  $\mathbf{\Lambda}_{k+1} = \mathbf{\Lambda}_k + \beta_k((\mathbf{X}\mathbf{U}_{k+1})\mathbf{\Sigma}_{k+1}\mathbf{V}_{k+1}^T + \mathbf{E}_{k+1} - \mathbf{X})$ .  
    **Step 4:** Update  $\beta_{k+1}$  by (11)-(12).  
    **Step 5:**  $k \leftarrow k + 1$ .  
**end while**

---

The linearization technique has also been applied to other optimization methods. For example, Yin [23] applied this technique to the Bregman iteration for solving compressive sensing problems and proved that the linearized Bregman method converges to an exact solution *conditionally*. In comparison, LADMAP always converges to an exact solution.

## 4 Experimental Results

In this section, we report numerical results on the standard LADMAP, the accelerated LADMAP and other state-of-the-art algorithms, including APG, ADM<sup>5</sup> and LADM, for LRR based data clustering problems. APG, ADM, LADM and LADMAP all utilize the Matlab version of PROPACK [11]. For the accelerated LADMAP, we provide two function handles to PROPACK which fulfils the successive matrix-vector multiplications. All experiments are run and timed on a PC with an Intel Core i5 CPU at 2.67GHz and with 4GB of memory, running Windows 7 and Matlab version 7.10.

We test and compare these solvers on both synthetic multiple subspaces data and the real world motion data (Hopkin155 motion segmentation database [18]). For APG, we set the parameters  $\beta_0 = 0.01$ ,  $\beta_{\min} = 10^{-10}$ ,  $\theta = 0.9$  and the Lipschitz constant  $\tau = \sigma_{\max}^2(\mathbf{X})$ . The parameters of ADM and LADM are the same as that in [13] and [21], respectively. In particular, for LADM the penalty is fixed at  $\beta = 2.5/\min(m, n)$ , where  $m \times n$  is the size of  $\mathbf{X}$ . For LADMAP, we set  $\varepsilon_1 = 10^{-4}$ ,  $\varepsilon_2 = 10^{-5}$ ,  $\beta_0 = \min(m, n)\varepsilon_2$ ,  $\beta_{\max} = 10^{10}$ ,  $\rho_0 = 1.9$ , and  $\eta_X = 1.02\sigma_{\max}^2(\mathbf{X})$ . As the code of ADM was downloaded, its stopping criteria,  $\|\mathbf{X}\mathbf{Z}_k + \mathbf{E}_k - \mathbf{X}\|/\|\mathbf{X}\| \leq \varepsilon_1$  and  $\max(\|\mathbf{E}_k - \mathbf{E}_{k-1}\|/\|\mathbf{X}\|, \|\mathbf{Z}_k - \mathbf{Z}_{k-1}\|/\|\mathbf{X}\|) \leq \varepsilon_2$ , are used in all our experiments<sup>6</sup>.

### 4.1 On Synthetic Data

The synthetic test data, parameterized as  $(s, p, d, \tilde{r})$ , is created by the same procedure in [13].  $s$  independent subspaces  $\{\mathcal{S}_i\}_{i=1}^s$  are constructed, whose bases  $\{\mathbf{U}_i\}_{i=1}^s$  are generated by  $\mathbf{U}_{i+1} = \mathbf{T}\mathbf{U}_i$ ,  $1 \leq i \leq s-1$ , where  $\mathbf{T}$  is a random rotation and  $\mathbf{U}_1$  is a  $d \times \tilde{r}$  random orthogonal matrix. So each subspace has a rank of  $\tilde{r}$  and the data has an ambient dimension of  $d$ . Then  $p$  data points are sampled from each subspace by  $\mathbf{X}_i = \mathbf{U}_i\mathbf{Q}_i$ ,  $1 \leq i \leq s$ , with  $\mathbf{Q}_i$  being an  $\tilde{r} \times p$  i.i.d. zero mean unit variance Gaussian matrix  $\mathcal{N}(0, 1)$ . 20% samples are randomly chosen to be corrupted by adding Gaussian noise with zero mean and standard deviation  $0.1\|\mathbf{x}\|$ . We empirically find that LRR achieves the best clustering performance on this data set when  $\mu = 0.1$ . So we test all algorithms with  $\mu = 0.1$  in this experiment. To measure the relative errors in the solutions, we run the standard LADMAP 2000 iterations with  $\beta_{\max} = 10^3$  to establish the ground truth solution  $(\mathbf{E}_0, \mathbf{Z}_0)$ .

---

<sup>5</sup>We use the Matlab code provided online by the authors of [13].

<sup>6</sup>Note that the second criterion differs from that in (19). However, this does not harm the convergence of LADMAP because (19) is always checked when updating  $\beta_{k+1}$  (see (12)).

The computational comparison is summarized in Table 1. We can see that the iteration numbers and the CPU times of both the standard and accelerated LADMAP are much less than those of other methods, and the accelerated LADMAP is further much faster than the standard LADMAP. Moreover, the advantage of the accelerated LADMAP is even greater when the ratio  $\tilde{r}/p$ , which is roughly the ratio of the rank of  $\mathbf{Z}_0$  to the size of  $\mathbf{Z}_0$ , is smaller, which testifies to the complexity estimations on the standard and accelerated LADMAP for LRR. It is noteworthy that the iteration numbers of ADM and LADM seem to grow with the problem sizes, while that of LADMAP is rather constant. Moreover, LADM is not faster than ADM. In particular, on the last data we were unable to wait until LADM stopped. Finally, as APG converges to an approximate solution to (2), its relative errors are larger and its clustering accuracy is lower than ADM and LADM based methods.

Table 1: Comparison among APG, ADM, LADM, standard LADMAP and accelerated LADMAP (denoted as LADMAP(A)) on the synthetic data. For each quadruple  $(s, p, d, \tilde{r})$ , the LRR problem, with  $\mu = 0.1$ , was solved for the same data using different algorithms. We present typical running time (in  $\times 10^3$  seconds), iteration number, relative error (%) of output solution  $(\hat{\mathbf{E}}, \hat{\mathbf{Z}})$  and the clustering accuracy (%) of tested algorithms, respectively.

Size $(s, p, d, \tilde{r})$	Method	Time	Iter.	$\frac{\ \hat{\mathbf{Z}} - \mathbf{Z}_0\ }{\ \mathbf{Z}_0\ }$	$\frac{\ \hat{\mathbf{E}} - \mathbf{E}_0\ }{\ \mathbf{E}_0\ }$	Acc.
(10, 20, 200, 5)	APG	0.0332	110	2.2079	1.5096	81.5
	ADM	0.0529	176	0.5491	0.5093	<b>90.0</b>
	LADM	0.0603	194	<b>0.5480</b>	<b>0.5024</b>	<b>90.0</b>
	LADMAP	0.0145	<b>46</b>	<b>0.5480</b>	<b>0.5024</b>	<b>90.0</b>
	LADMAP(A)	<b>0.0010</b>	<b>46</b>	<b>0.5480</b>	<b>0.5024</b>	<b>90.0</b>
(15, 20, 300, 5)	APG	0.0869	106	2.4824	1.0341	80.0
	ADM	0.1526	185	0.6519	0.4078	83.7
	LADM	0.2943	363	<b>0.6518</b>	<b>0.4076</b>	<b>86.7</b>
	LADMAP	0.0336	<b>41</b>	<b>0.6518</b>	<b>0.4076</b>	<b>86.7</b>
	LADMAP(A)	<b>0.0015</b>	<b>41</b>	<b>0.6518</b>	<b>0.4076</b>	<b>86.7</b>
(20, 25, 500, 5)	APG	1.8837	117	2.8905	2.4017	72.4
	ADM	3.7139	225	1.1191	1.0170	80.0
	LADM	8.1574	508	<b>0.6379</b>	<b>0.4268</b>	80.0
	LADMAP	0.7762	<b>40</b>	<b>0.6379</b>	<b>0.4268</b>	<b>84.6</b>
	LADMAP(A)	<b>0.0053</b>	<b>40</b>	<b>0.6379</b>	<b>0.4268</b>	<b>84.6</b>
(30, 30, 900, 5)	APG	6.1252	116	3.0667	0.9199	69.4
	ADM	11.7185	220	0.6865	0.4866	<b>76.0</b>
	LADM	N.A.	N.A.	N.A.	N.A.	N.A.
	LADMAP	2.3891	<b>44</b>	<b>0.6864</b>	<b>0.4294</b>	<b>80.1</b>
	LADMAP(A)	<b>0.0058</b>	<b>44</b>	<b>0.6864</b>	<b>0.4294</b>	<b>80.1</b>

Table 2: Comparison among APG, ADM, LADM, standard LADMAP and accelerated LADMAP on the Hopkins155 database. We present their average computing time (in seconds), average number of iterations and average classification errors (%) on all 156 sequences.

	Two Motion			Three Motion			All		
	Time	Iter.	CErr.	Time	Iter.	CErr.	Time	Iter.	CErr.
APG	15.7836	90	5.77	46.4970	90	<b>16.52</b>	22.6277	90	8.36
ADM	53.3470	281	<b>5.72</b>	159.8644	284	<b>16.52</b>	77.0864	282	<b>8.33</b>
LADM	9.6701	110	5.77	22.1467	64	<b>16.52</b>	12.4520	99	8.36
LADMAP	3.6964	<b>22</b>	<b>5.72</b>	10.9438	<b>22</b>	<b>16.52</b>	5.3114	<b>22</b>	<b>8.33</b>
LADMAP(A)	<b>2.1348</b>	<b>22</b>	<b>5.72</b>	<b>6.1098</b>	<b>22</b>	<b>16.52</b>	<b>3.0202</b>	<b>22</b>	<b>8.33</b>

## 4.2 On Real World Data

We further test the performance of these algorithms on the Hopkins155 database [18]. This database consists of 156 sequences, each of which has 39 to 550 data vectors drawn from two or three motions. For computational efficiency, we preprocess the data by projecting it to be 5-dimensional using PCA. As  $\mu = 2.4$  is the best parameter for this database [13], we test all algorithms with  $\mu = 2.4$ .

Table 2 shows the comparison among APG, ADM, LADM, standard LADMAP and accelerated LADMAP on this database. We can also see that the standard and accelerated LADMAP are much faster than APG, ADM, and LADM, and the accelerated LADMAP is also faster than the standard LADMAP. However, in this experiment the advantage of the accelerated LADMAP over the standard LADMAP is not as dramatic as that in Table 1. This is because on this data  $\mu$  is chosen as 2.4, which cannot make the rank of the ground truth solution  $\mathbf{Z}_0$  much smaller than the size of  $\mathbf{Z}_0$ .

## 5 Conclusions

In this paper, we propose a linearized alternating direction method with adaptive penalty (LADMAP) and apply it to solving the LRR problem. With linearization, auxiliary variables need not be introduced for closed-form solutions, when the objective functions are matrix or vector norms. Moreover, with fewer variables and constraints, the convergence also becomes faster. Allowing the penalty to change adaptively further accelerates the convergence of LADM. When applying LADMAP to LRR, by representing the representation matrix as its skinny SVD, full sized matrix-matrix multiplications are avoided by using successive reduced sized matrix-matrix multiplications instead, and successive matrix-vector multiplications are introduced to compute the partial SVD. Finally, we are able to solve LRR at a computation complexity of  $O(rn^2)$ , which is highly advantageous over the existing LRR solvers. Numerical results demonstrate that LADMAP converges faster than LADM and ADM and our acceleration techniques are effective on LRR. Although we only present results on LRR, LADMAP is actually a general method that can be applied to other convex programs. We will test it with more problems in sparse representation in the future.

## A Proof of Theorem 3

**Proof** By Proposition 2 (1),  $\{(\mathbf{x}_k, \mathbf{y}_k, \lambda_k)\}$  is bounded, hence has an accumulation point, say  $(\mathbf{x}_{k_j}, \mathbf{y}_{k_j}, \lambda_{k_j}) \rightarrow (\mathbf{x}^\infty, \mathbf{y}^\infty, \lambda^\infty)$ . We accomplish the proof in two steps.

1. We first prove that  $(\mathbf{x}^\infty, \mathbf{y}^\infty, \lambda^\infty)$  is a KKT point of problem (1).

By Proposition 2 (2),  $\mathcal{A}(\mathbf{x}_{k+1}) + \mathcal{B}(\mathbf{y}_{k+1}) - \mathbf{c} = \beta_k^{-1}(\lambda_{k+1} - \lambda_k) \rightarrow 0$ . This shows that any accumulation point of  $\{(\mathbf{x}_k, \mathbf{y}_k)\}$  is a feasible solution.

By letting  $k = k_j - 1$  in Proposition 1 and the definition of subgradient, we have

$$\begin{aligned} f(\mathbf{x}_{k_j}) + g(\mathbf{y}_{k_j}) &\leq f(\mathbf{x}^*) + g(\mathbf{y}^*) + \langle \mathbf{x}_{k_j} - \mathbf{x}^*, -\beta_{k_j-1}\eta_A(\mathbf{x}_{k_j} - \mathbf{x}_{k_j-1}) - \mathcal{A}^*(\tilde{\lambda}_{k_j}) \rangle \\ &\quad + \langle \mathbf{y}_{k_j} - \mathbf{y}^*, -\beta_{k_j-1}\eta_B(\mathbf{y}_{k_j} - \mathbf{y}_{k_j-1}) - \mathcal{B}^*(\hat{\lambda}_{k_j}) \rangle. \end{aligned}$$

Let  $j \rightarrow +\infty$ , by observing Proposition 2 (2), we have

$$\begin{aligned} f(\mathbf{x}^\infty) + g(\mathbf{y}^\infty) &\leq f(\mathbf{x}^*) + g(\mathbf{y}^*) + \langle \mathbf{x}^\infty - \mathbf{x}^*, -\mathcal{A}^*(\lambda^\infty) \rangle + \langle \mathbf{y}^\infty - \mathbf{y}^*, -\mathcal{B}^*(\lambda^\infty) \rangle \\ &= f(\mathbf{x}^*) + g(\mathbf{y}^*) - \langle \mathcal{A}(\mathbf{x}^\infty - \mathbf{x}^*), \lambda^\infty \rangle - \langle \mathcal{B}(\mathbf{y}^\infty - \mathbf{y}^*), \lambda^\infty \rangle \\ &= f(\mathbf{x}^*) + g(\mathbf{y}^*) - \langle \mathcal{A}(\mathbf{x}^\infty) + \mathcal{B}(\mathbf{y}^\infty) - \mathcal{A}(\mathbf{x}^*) - \mathcal{B}(\mathbf{y}^*), \lambda^\infty \rangle \\ &= f(\mathbf{x}^*) + g(\mathbf{y}^*), \end{aligned}$$

where we have used the fact that both  $(\mathbf{x}^\infty, \mathbf{y}^\infty)$  and  $(\mathbf{x}^*, \mathbf{y}^*)$  are feasible solutions. So we conclude that  $(\mathbf{x}^\infty, \mathbf{y}^\infty)$  is an optimal solution to (1).

Again, let  $k = k_j - 1$  in Proposition 1 and by the definition of subgradient, we have

$$f(\mathbf{x}) \geq f(\mathbf{x}_{k_j}) + \langle \mathbf{x} - \mathbf{x}_{k_j}, -\beta_{k_j-1}\eta_A(\mathbf{x}_{k_j} - \mathbf{x}_{k_j-1}) - \mathcal{A}^*(\tilde{\lambda}_{k_j}) \rangle, \quad \forall \mathbf{x}. \quad (20)$$

Fix  $\mathbf{x}$  and let  $j \rightarrow +\infty$ , we see that

$$f(\mathbf{x}) \geq f(\mathbf{x}^\infty) + \langle \mathbf{x} - \mathbf{x}^\infty, -\mathcal{A}^*(\lambda^\infty) \rangle, \quad \forall \mathbf{x}.$$

So  $-\mathcal{A}^*(\lambda^\infty) \in \partial f(\mathbf{x}^\infty)$ . Similarly,  $-\mathcal{B}^*(\lambda^\infty) \in \partial g(\mathbf{y}^\infty)$ . Therefore,  $(\mathbf{x}^\infty, \mathbf{y}^\infty, \lambda^\infty)$  is a KKT point of problem (1).

2. We next prove that the whole sequence  $\{(\mathbf{x}_k, \mathbf{y}_k, \lambda_k)\}$  converges to  $(\mathbf{x}^\infty, \mathbf{y}^\infty, \lambda^\infty)$ .

By choosing  $(\mathbf{x}^*, \mathbf{y}^*, \lambda^*) = (\mathbf{x}^\infty, \mathbf{y}^\infty, \lambda^\infty)$  in Proposition 2, we have  $\eta_A \|\mathbf{x}_{k_j} - \mathbf{x}^\infty\|^2 - \|\mathcal{A}(\mathbf{x}_{k_j} - \mathbf{x}^\infty)\|^2 + \eta_B \|\mathbf{y}_{k_j} - \mathbf{y}^\infty\|^2 + \beta_{k_j}^{-2} \|\lambda_{k_j} - \lambda^\infty\|^2 \rightarrow 0$ . By Proposition 2 (1), we readily have



$\eta_A \|\mathbf{x}_k - \mathbf{x}^\infty\|^2 - \|\mathcal{A}(\mathbf{x}_k - \mathbf{x}^\infty)\|^2 + \eta_B \|\mathbf{y}_k - \mathbf{y}^\infty\|^2 + \beta_k^{-2} \|\lambda_k - \lambda^\infty\|^2 \rightarrow 0$ . So  $(\mathbf{x}_k, \mathbf{y}_k, \lambda_k) \rightarrow (\mathbf{x}^\infty, \mathbf{y}^\infty, \lambda^\infty)$ .

As  $(\mathbf{x}^\infty, \mathbf{y}^\infty, \lambda^\infty)$  can be an arbitrary accumulation point of  $\{(\mathbf{x}_k, \mathbf{y}_k, \lambda_k)\}$ , we may conclude that  $\{(\mathbf{x}_k, \mathbf{y}_k, \lambda_k)\}$  converges to a KKT point of problem (1).

## References

- [1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. In Michael Jordan, editor, *Foundations and Trends in Machine Learning*, 2010.
- [2] J. Cai, E. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *preprint*, 2008.
- [3] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 2011.
- [4] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 2009.
- [5] E. J. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 2008.
- [6] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *CVPR*, 2009.
- [7] P. Favaro, R. Vidal, and A. Ravichandran. A closed form solution to robust subspace estimation and clustering. In *CVPR*, 2011.
- [8] A. Goldberg, X. Zhu, B. Recht, J. Xu, and R. Nowak. Transduction with matrix completion: Three birds with one stone. In *NIPS*, 2010.
- [9] B.S. He, H. Yang, and S.L. Wang. Alternating direction method with self-adaptive penalty parameters for monotone variational inequality. *J. Optimization Theory and Applications*, 106:337–356, 2000.
- [10] H. Ji, C. Liu, Z. Shen, and Y. Xu. Robust video denoising using low rank matrix completion. In *CVPR*, 2010.
- [11] R. Larsen. Lanczos bidiagonalization with partial reorthogonalization. *Department of Computer Science, Aarhus University, Technical report, DAIMI PB-357*, 1998.
- [12] Z. Lin, M. Chen, L. Wu, and Y. Ma. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. *UIUC Technical Report UILU-ENG-09-2215, arxiv:1009.5055*, 2009.
- [13] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, 2010.
- [14] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient  $l_{2,1}$  norm minimization. In *UAI*, 2009.
- [15] Y. Ni, J. Sun, X. Yuan, S. Yan, and L. Cheong. Robust low-rank subspace segmentation with semidefinite guarantees. In *ICDM Workshop*, 2010.
- [16] M. Tao and X.M. Yuan. Recovering low-rank and sparse components of matrices from incomplete and noisy observations. *SIAM Journal on Optimization*, 21(1):57–81, 2011.
- [17] K. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *preprint*, 2009.
- [18] R. Tron and R. Vidal. A benchmark for the comparison of 3D motion segmentation algorithms. In *CVPR*, 2007.
- [19] J. Wright, A. Ganesh, S. Rao, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *NIPS*, 2009.
- [20] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 2010.
- [21] J. Yang and X. Yuan. Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. *submitted*, 2011.

486 [22] J. Yang and Y. Zhang. Alternating direction algorithms for  $l_1$  problems in compressive sensing.  
487 *SIAM J. Scientific Computing*, 2010.  
488  
489 [23] W. Yin. Analysis and generalizations of the linearized Bregman method. *SIAM Journal on*  
490 *Imaging Sciences*, 2010.  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539